BY DEREK RILEY

How Large Language Models Can Assist Litigation Workflows

Law firms that effectively embrace large language model (LLM)-based technology can gain a significant competitive advantage particularly in high-volume or high-analysis workflows such as contract drafting, regulatory work, and numerous aspects of litigation, including legal research, briefing, and trial preparation. This article provides a basic introduction to LLMs for legal professionals.

Foundation large language models (LLMs) – such as OpenAI's GPT-5, Google Gemini, Meta's LLaMA, and Anthropic's Claude – are rapidly accelerating workflows across many industries due to their unique and flexible capabilities to understand content and generate meaningful text. The foundation models are trained using publicly available text to generate output that is broadly useful. The capabilities of these models with a wide variety of tasks have led to their usage in specific industries, including law. However, foundation models have inherent limitations that can lead to the generation of text that looks accurate but contains semantic flaws, often called hallucinations. Moreover, without a paid version of the foundation LLMs, or other programming, the models will train on the data they receive from the user. Accordingly, they must be used with caution.

Lawyers with expertise with language, meaning, and content are uniquely suited to leverage LLMs. Rather than replacing attorneys, LLMs have the capacity to augment legal expertise, streamline operations, and reduce drudge work. Law firms that are effectively embracing this technology can gain a significant competitive advantage particularly in high-volume or high-analysis workflows such as contract drafting, regulatory work, and numerous aspects of litigation, including legal research, briefing, and trial preparation. In the future, it might be hard to imagine a client willing to pay for the time an attorney spends to perform tasks without the benefit of AI any more than a customer would pay a tailor to make clothes by hand.

This article explains how foundation LLMs work, how they can be augmented with legal-case-specific data to provide security and reduce hallucinations through Retrieval Augmented Generation (RAG) architectures, and the fundamentals of creating prompts to retrieve the desired output. Further, the article examines the current marketplace of AI tools and how lawyers can differentiate between them, ethical and professional obligations in their adoption, and how lawyers can expect LLMs to affect the business and practice of law. (See the related content in this issue of *Wisconsin Lawyer*: Timothy D. Edwards, *Challenging the Foundations of 'Deepfake' Evidence*; and Bonnie J. Shucha, *NotebookLM for Lawyers: AI That Focuses on Your Documents*.)

How LLMs Work: A Technical Overview for Legal Professionals

Generative Pre-trained Transformers. Foundation LLMs are built upon a specific architecture of artificial neural networks called the transformer architecture.[1] Successful training of LLMs has been enabled by recent developments in computing capabilities by NVIDIA (a technology company headquartered in Santa Clara, Cal.) to train on extremely large data sets. Foundation models have been trained on essentially all available text on the internet (trillions of words). It would take a human 11 million years to read the same amount of text. Modern training approaches allow new foundation models to be trained on this data in a matter of days by using supercomputers in which the large computational task of training is broken into smaller parts and executed simultaneously with numerous graphics processing units (GPUs), each with thousands of specialized cores that can perform many calculations in parallel.

The models are trained to perform a wide range of language-based tasks, from summarization and translation to legal analysis and argumentation. Ultimately, LLMs are next-token prediction models, so they generate essentially one word at a time and are able to produce large amounts of syntactically correct responses with semantic validity, if they have appropriate context. Lawyers who know how to effectively provide the appropriate context– potentially through other software – will be able to leverage these tools to maximum utility for their clients.

Training on Massive Datasets. LLM foundation models are "pre-trained" on diverse text including legal cases, governmental filings, websites, books, and other digital documents. The data that is used to train models is internalized in the model, and it helps establish the core patterns that the model follows. Foundation LLMs possess relevant domain knowledge for many legal tasks, but the application of such knowledge can lead to responses that contain hallucinations. This is akin to law students who have studied every case but struggle to apply their knowledge to situations that they lack the context to understand fully.

To overcome the problem of hallucinations, critical domain information can be added to the prompt in the "context window" to give the model a concrete foundation for its responses. This is similar to giving a law student key documents and providing some guiding principles on what is expected in the summary. While this approach of interacting with an LLM may seem obvious, the specific nuances for which context to provide, terms to use in the prompt, and which foundation LLM to use can lead to widely varying results in output, which often leads to discontent with foundation models.[2]

The Black Box Problem. While outputs from LLMs can be highly syntactically or semantically accurate or both, the internal mechanisms of the LLM that were used to lead to the output remain largely opaque – a phenomenon referred to as the "black box" problem.[3] The model's decisions result from complex internal representations that are difficult to explain in human terms, and while LLMs can be asked to justify their output, there are no guarantees that the justifications are representative of the underlying decisions that the LLM uses.

Confidence in the quality of output is often gained through using the systems over time, but care must always be taken to check the output to validate its accuracy. Empirical evaluations have shown LLMs can outperform new attorneys on multiple-choice legal-reasoning tasks and contract-drafting assessments.[4] This is possible due to the quantity of data these models have contextualized as well as the capacity of the models. Still, confidence with a model should never reach the point of complacency akin to working with a trusted colleague over time. Ideally, software solutions will be built for workflows in which citation can be readily accomplished or make the process quicker for more complex legal analysis.

Context Window. All foundation LLMs have a limited context window, which determines the amount of text it can ingest in a single session. GPT-4 Turbo, for example, supports up to 128,000 tokens, which is approximately the length of a 300-page novel. However, LLMs might not pay attention to everything in the context window. Research has shown that LLMs tend to ignore context in the middle of a long context window, preferring to pay attention to context toward the beginning and end of the window.[5] Deciding which text is included in the input and how it is organized is an ongoing area of research.

Context-window limitations have significant implications for uploading case files, deposition transcripts, or statutes in full. If the input is too long, it might be cut off at the beginning or end, leading to missing key information or inaccurate responses. In addition, with full legal filings or a long transcript, the model might focus on the most recent or prominent parts or ignore context spread throughout the document. Foundation model interfaces such as ChatGPT do not give users the ability to control all aspects of the context window, so inaccuracies and hallucinations are common occurrences.[6]

Integrating Legal-Case-Specific Data with LLMs

General-purpose foundation LLMs excel in generating well-reasoned arguments and summaries but lack familiarity with domain-specific legal content. While some organizations initially sought to fine-tune their own LLMs on proprietary data, this has been resource intensive and largely redundant with advances in the foundation LLMs themselves. Instead, most tech-forward firms now use software featuring a RAG process to identify and manage domain-specific context without retraining the foundation model.

What Is RAG? RAG is a technique that combines LLMs with a specialized, high-performance context database. When a user inputs a query, the system searches the databases for relevant documents (for example, contracts, filings, depositions), which are then passed to the LLM in the context window to generate a grounded response. This approach avoids training on sensitive data, thereby ensuring security compliance. There are many established RAG architectures that can be engineered to address the complexity of different tasks in the legal workflow.

Derek Riley, Ph.D., is a professor and the director of the Milwaukee School of Engineering's (MSOE) computer science program, which includes an emphasis in artificial intelligence (AI) and deep learning. His teaching and research areas include machine learning, modeling, and high-performance computing. Riley earned his Ph.D. in 2009 at Vanderbilt University, where he developed high-performance formal modeling and simulation methods for biochemical and industrial systems. In addition to teaching at MSOE, he provides data, algorithm, large language model, and AI consulting services, and is a member of the Association for Computing Machinery.

He thanks Atty. Michael Aiken for his contributions to this article. Access the digital article at www.wisbar.org/wl.

riley@msoe.edu

Model Plug-and-Play. Because RAG separates the knowledge base from the language model, newer or better LLMs can be swapped into existing pipelines without redesigning the system. This modularity ensures legal organizations stay up to date with the latest models while maintaining their proprietary knowledge infrastructure. Benchmarking of foundation models on legal reasoning tasks is a standard practice when new models are released due to the commonality of this use case.[7]

Safely Incorporating LLMs into Legal Workflows

Properly prompted and contextualized LLMs can generate cited, verifiable outputs that make the human validation process efficient and consistent. While RAG appears to improve the performance of language models in answering legal queries, the hallucination problem can still persist.[8] As with obtaining information from an article online, lawyers must never rely solely on AI outputs and must always validate the response with a detailed manual review. Courts have sanctioned attorneys numerous times for submitting filings based on AI-generated content without verification.[9] This human-in-the-loop approach ensures ethical compliance, minimizes risk, and maintains accountability.[10] Software applications can be designed to make checking AI citations more efficient than paper files.

Lawyers using LLM technology must adhere to both ethical obligations and practical concerns.[11] Wisconsin attorneys are bound by the Rules of Professional Conduct for Attorneys, codified in SCR chapter 20, including SCR 20:1.1 (Competence), which now includes technological competence, and SCR 20:1.4 (Communication). Lawyers must understand "the benefits and risks associated with relevant technology" to comply with these rules and communicate their intentions with regard to their use of such technology when discussing the scope of the representation with clients. Licensing terms for foundation models hosted by cloud-based AI services stipulate whether data is used for training and should be reviewed to ensure there is no potential breach of the attorney-client privilege. Generally free-tier models do *not* guarantee that data will be kept private, so paid subscriptions are necessary, but not sufficient to ensure privacy is preserved.

Prompt Engineering

Aside from the RAG architecture, the quality of LLM output depends heavily on the quality of the prompt.[12] "Prompt engineering" refers to the structured development of instructions given to an LLM to guide it toward a desired outcome. Generally, a user should be maximally prescriptive because word choice, style, tone, structure, and context of the prompt all matter. In addition, it is often recommended to assign the LLM an explicit role. There is no need to be polite; the user should get straight to the point in uncertain terms. Results are likely to be better when users provide affirmative directives such as "do" rather than negative language such as "don't" and repeat a specific word or phrase as needed to focus the inquiry.[13]

Effective prompts are **specific**: they clearly state the task; **context rich**: they provide background information, tone, and format; and **structured**: they explicitly define the expected output and audience.

For example, this is an effective prompt:

Task: Identify all the evidence to support a breach of fiduciary duty claim against Mr. Smith.

Context: You are a litigation assistant specializing in breach of fiduciary duty cases and focus on contradictions between witnesses, admissions, or credibility issues that relate to the alleged self-dealing and concealment of material facts.

Output: Create a list of each piece of evidence and explain how it supports a claimed breach step by step, with citations to the page and line of any relevant testimony or exhibits for reference.

Researchers recommend zero-shot (no examples), few-shot (two to six examples), and chain-of-thought prompting (ask for step-by-step reasoning or going through the steps with individual questions), depending on task complexity.[14] Techniques such as **step-back prompting**, in which the user asks the LLM to consider a more general knowledge question first to activate relevant background information, and then consider a specific application; **automated prompting**, in which the LLM itself is used to generate prompts; and **prompt self-evaluation**, asking the model what it found unclear, are also emerging best practices. Prompting is a naturally iterative process. Some software includes prompt histories and prompt libraries for common tasks.

LLMs in the Legal Software Marketplace

A rapidly growing marketplace of legal AI tools is leveraging foundation LLMs to offer specialized capabilities. (The listed features are from the companies' websites; no endorsement of these products is intended.)

- **Alexi.com** – AI powered legal research, document summarization, and litigation prep.
- **August.law** – A configurable legal AI platform targeted to mid-size firms for creating modular AI agents for firm specific workflows.
- **Briefpoint.ai** – Automated written discovery drafting.
- **Strongsuit.com –** AI-powered legal research and drafting.
- **Clio.com** – AI-integrated practice management software.
- **Disco's Celia** – A suite of generative AI tools designed to streamline discovery workflows.
- **Gavel.io** – Microsoft Word add-in for AI-powered redlining, drafting, and negotiation, which uses multiple foundation models to address corporate and real estate transactions.
- **Iqidis.ai** – Personalized RAG system to perform legal research, drafting, and analysis with inspection of citation sources.
- **Filevine.com** – AI-enhanced case management, document information extraction, organization, deposition analysis, and calendaring. Promises to extract key information with data mapping and generate demand letters in minutes with supporting facts.
- **Skribe.ai / Depo Copilot** – Deposition recording, summarization, and analysis.
- **Harvey.ai** – End-to-end legal assistant built on OpenAI. High-volume document analysis to extract key information and summarize. Promises specialization in legal brief writing and Outlook integration.
- **Lawme.ai** – Suite of AI-powered tools on client onboarding, bulk data extraction, legal research, and contract drafting.
- **LawLM.ai** – Focus on AI-generated deposition summaries and analysis with a chatbot to analyze testimonial evidence across multiple witnesses. Offers no-subscription product for collaboration between lawyers, experts, and clients.
- **Lexis+** – Lexis Nexis proprietary LLM and tool suite. Also features AI-assisted case law research with Lexis content.
- **Matey.ai** – Secure and scalable AI workflows related to document analysis and trial readiness.
- **TryNovo.com** – Tools for demand letters and medical chronologies.
- **Paxton.ai** – AI-powered legal research and drafting platform, designed

to streamline tedious tasks and enhance productivity, with confidence indicator and AI citator.

- **SmartAdvocate.com** – Case management with AI tools to summarize cases, briefs, and records and with other integrated non-AI features.
- **Supio.com** – Specialized AI for plaintiff's personal injury document formatting and data.
- **Spellbook.legal** – GPT-4 contract drafting and review tool.
- **Cocounsel Legal** – Thomson Reuters LLM tool suite that includes analysis of large sets of legal documents and extraction of key documents and information, along with document drafting. Also features AI-assisted case law research with Westlaw content.

These tools span document routing and coding, legal research, discovery, case management, evidence summary and review, trial preparation, and client engagement. They promise to democratize legal access by enhancing a firm's productive capabilities, allowing lawyers to focus on high-value strategic work, and reducing costs for clients.

Not all the tools are created equal. The architecture of the software critically influences RAG performance and the legal usefulness of responses.[15] Some are little more than "ChatGPT wrappers" with minimal added value aside from some guardrails and user-interface changes, without lawyers involved in development. Lawyers should look for features such as custom databases, RAG integration, citation management, and workflow to distinguish robust and effective AI solutions from superficial ones. Many AI thought leaders believe these platforms will lead to increased ease of collaboration between lawyers, clients, and experts. Consultant firms and comparison websites are cropping up to help advise firms on finding the right approach and fit for their specific needs, but the best way to find the right fit is to try products. Some products have nonpublic subscription pricing and closed demonstrations, while others offer transparent pricing and open, free trials.

Properly prompted LLMs can generate cited, verifiable outputs that make the validation process efficient and consistent.

The Future of LLMs in Legal Practice

Generative AI is still in the early stages of disrupting the legal industry. As LLMs become more capable, secure, and explainable, they will continue to be adopted by more firms. Large firms may attempt to build proprietary models on internal data as a competitive advantage. Small to mid-size firms may favor the economics of leveraging general-purpose LLMs via custom RAG pipelines to obtain the bandwidth to compete with big-firm resources. The new tools can help to craft more persuasive narratives and help with objectivity in evaluating case evidence, resulting in better advocacy and more efficient resolutions.

They will also likely cause more people who are not lawyers to feel more comfortable representing themselves in court. Ethical considerations, billing-model adjustments (for example, reduced reliance on the billable hour), lawyer psychology,[16] and bar association guidance will shape this evolution. Although financial incentives are lacking for lawyers to adopt technology that could mean a short term reduction in billable revenue, firms that fail to embrace LLMs may struggle to remain competitive in the long term. **WL**

ENDNOTES

[1] Ashish Vaswani et al., *Attention Is All You Need*, in Advances in Neural Information Processing Systems 30 (NIPS 2017), https://arxiv.org/abs/1706.03762.

[2] *See, e.g.,* Bonnie J. Shucha, *AI Prompting for Legal Professionals: The Art of Asking the Right Question*, 98 Wis. Law. 29 (Nov. 2025).

[3] Zachary C. Lipton, *The Mythos of Model Interpretability*, arXiv (June 2016), https://arxiv.org/abs/1606.03490.

[4] Michael J. Bommarito II & Daniel Martin Katz, *GPT Takes the Bar Exam*, SSRN Elec. J. (2023), https://doi.org/10.2139/ssrn.4314839.

[5] Nelson F. Liu et al., *Lost in the Middle: How Language Models Use Long Contexts*, arXiv (July 2023), https://arxiv.org/abs/2307.03172.

[6] Varun Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, J. of Empirical Legal Stud., 2025; 0:1-27, https://arxiv.org/abs/2405.20362.

[7] Yu Fan et al., *LEXam: Benchmarking Legal Reasoning on 340 Law Exams*, arXiv (May 2025), https://arxiv.org/abs/2505.12864; also available at SSRN, https://ssrn.com/abstract=5265144.

[8] Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, in Advances in Neural Information Processing Systems 33 (2020).

[9] *See, e.g.*, Sara Merken, *Lawyers in Walmart Lawsuit Admit AI Hallucinated Case Citations*, Reuters (Feb. 10, 2025), https://www.reuters.com/legal/legalindustry/lawyers-walmart-lawsuit-admit-ai-hallucinated-case-citations-2025-02-10/.

[10] Lee Boonstra, *Prompt Engineering* (Google White Paper, Sept. 2024), https://www.gptaiflow.com/assets/files/2025-01-18-pdf-1-TechAI-Goolge-whitepaper_Prompt%20Engineering_v4-af36dcc7a49bb7269a58b1c9b89a8ae1.pdf.

[11] See ABA Comm. on Ethics & Prof'l Responsibility, Formal Op. 512 (2023), https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility /ethics-opinions/aba-formal-opinion-512.pdf.

[12] Boonstra, *supra* note 10.

[13] *See* Shucha, *supra* note 2.

[14] Boonstra, *supra* note 10; Jason Wei et al., *Chain of Thought Prompting Elicits Reasoning in Large Language Models*, arXiv (Jan. 2022), https://arxiv.org/abs/2201.11903.

[15] *See supra* note 11.

[16] Tom Martin, *AI is Like Ozempic: How the Shame of Using AI in Law Practice Mirrors Anxieties About Drug-Aided Weight Loss*, Lawdroid Manifesto (Aug. 14, 2025), https://www.lawdroidmanifesto.com/p/ai-is-like-ozempic-how-the-shame. **WL**